Biophysics, Vol. 44, No. 4, 1999, pp. 584–593. Translated from Biofizika, Vol. 44, No. 4, 1999, pp. 601–610.

MOLECULAR BIOPHYSICS

Computer Analysis of Regulatory Patterns in Complete Bacterial Genomes. Translation Initiation of the Ribosomal Protein Operons

A. G. Vitreshchak[1], A. K. Bansal[2], I. I. Titov[3], and M. S. Gelfand[4]

[1]*Institute of Problems of Information Transmission, Russian Academy of Sciences, Moscow, Russia*
[2]*Kent State University, Kent, OH 44242, USA*
[3]*Institute of Cytology and Genetics, Russian Academy of Sciences, Siberian Branch, Novosibirsk, 630090 Russia*
[4]*State Scientific Center GosNIIGenetika, Moscow, 113545 Russia*

Received September 21, 1997

Abstract—Translation initiation signals are predicted for the ribosomal protein operons of *Haemophilus influenzae*. This process is regulated by formation of RNA secondary structures that are bound by one of the proteins from the regulated operons. In some cases these structures imitate the binding site for the protein in the ribosomal RNA. Prediction was done by comparison with the homologous operons of *Escherichia coli* and the analogous structures in rRNA supplemented by computation of the hairpin formation energy. This regulatory mechanism has been shown for the *L11*, *S10*, *S15*, *spc*, and α operons of *H. influenzae*, and possibly for the *S15* operon of *Helicobacter pylori*, *Bacillus subtilis* and *Mycoplasma genitalium*.

Key words: computer analysis, functional signals, translation initiation, ribosomal proteins, complete genomes, *Haemophilus influenzae*

In *Escherichia coli*, initiation of translation of the ribosomal protein operons is regulated by binding of one of the encoded proteins with a hairpin or pseudoknot upstream of one of the proximal genes of the operon [1, 2]. The binding stabilizes the secondary structure element. Since the latter usually covers the Shine–Dalgarno box and the start codon, the secondary structure interferes with the ribosome binding and the translation does not initialize. Since the downstream genes are often translationally coupled with the upstream genes (this is one of the mechanisms for maintenance of the correct expression levels [1, 3]), translation of all genes stops. Thus the expression of the ribosomal proteins is regulated by autorepression when the protein is present in excess.

In some cases these structures imitate the binding site for the protein in the ribosomal RNA [2]. It should be noted that the binding constants of the protein to the mRNA structure are typically much lower than the binding constants to the ribosomal RNA. Thus the expression stops only at considerable excess of the free protein not bound to the ribosome. Sometimes there exist alternative hairpins and pseudoknots. The functional significance of these structures is not always clear.

Such regulatory mechanisms were demonstrated for several ribosomal proteins of *Escherichia coli*. However, analogous mechanisms for other bacteria, including those with completely sequenced genomes, have not been studied. In this work we attempt to predict the regulatory patterns in the ribosomal protein operons of *Haemophilus influenzae*.

The analysis is based on the so-called comparative approach to prediction of the RNA secondary structure [4]. It is based on the observation that the functionally important secondary structure RNA elements are more conserved than the nucleotide sequence itself. Thus, simultaneous folding of two or more sequences yields a functionally meaningful

structure. This approach was first used in 1969 for prediction of the tRNA structure [5] and since then was successfully applied to analysis of many more spatial structures. It should be noted that the results obtained by comparative analysis are usually superior to the predictions obtained by minimization of the free energy [6]. Indeed, the latter is unstable from the computational point of view: even small changes of parameters (e.g., stacking energy) lead to substantial changes in the optimal structure. Besides, the existing algorithms do not allow formation of pseudoknots and do not account for tertiary interactions (e.g., [7]). It should be noted, however, that until now the comparative approach has been used mostly for analysis of structural RNAs, whereas here we consider regulatory regions in mRNA.

The analysis was done as follows. The program GOLDIE 2.0 [8] was used to scan the complete genome of *Haemophilus influenzae* [9] in order to find gene chains orthologous to the ribosomal protein genes of *Escherichia coli*. Since the order and direction of these genes coincide, it is highly probable that the operon structure also is conserved. Homologous operons were found also in some other genomes, in particular, *Helicobacter pylori* [10], *Bacillus subtilis* [11], and *Mycoplasma genitalium* [12]. Proteins were aligned in order to determine exactly the gene starts.

Then comparison with the known *E. coli* structures was used to predict the secondary structures responsible for the regulation of translation initiation. The energy of the obtained structures was computed using the programs described in [37], at 37°C, with parameters from [14].

This resulted in the candidate regulatory structures for the following operons.

Operon *spc*. It is regulated by binding of the protein *S8* encoded by the fifth gene of the operon, with a long hairpin situated at the 5′-end of the third gene, *rplE* (L5). The hairpin formation energy is –23.0 kcal/mol in *E. coli* and –15.5 kcal/mol in *H. influenzae*. The Shine–Dalgarno box in both genomes is in a single strand segment to the left (5′) of the hairpin. The hairpins have bulge loops similar to the recognition site for S8 in the 16S rRNA (Fig. 1 right, cf. [15, Fig. 2]). It should be noted that two alternative hairpins can be constructed (Fig. 1 left and center). Both hairpins are supported by the comparative analysis, including complementary mutations. The same structure exists in other enterobacteria [15]. The region forming the alternative structures is absolutely conserved. This structure does not form in *B. subtilis* [16].

Operon *S10*. The regulatory protein L4 is encoded the third gene of the operon. The expression is regulated in two levels. First, hairpin 5 (Fig. 2) is an attenuator that terminates transcription [17]. Second, hairpin 6 covers the Shine–Dalgarno box and the start codon of the first gene *rpsJ* (S10) [17, 18]. Both these structures are present in *E. coli*) (total energy –50.7 kcal/mol) and *H. influenzae* (–44.0 kcal/mol). However, only one of the versions of the regulatory structure initially suggested in [19, Fig. 7] is supported by the comparative analysis. Hairpin 6, responsible for regulation on the translation level, has been described as similar to a site in 23S rRNA [17]. However, the comparative analysis does not confirm this hypothesis: although the corresponding fragments of the 23S rRNA of *E. coli* and *H. influenzae* are practically identical, the position of the candidate analogous regions in the mRNAs is absolutely different.

Operon *L11*. This operon consists of two genes *rplK* and *rplA* encoding proteins L11 and L1 respectively, The regulatory region is shown in Fig. 3. The translation does not initialize if the hairpin preceding *rplK* is bound by L1 [20]. Some elements of the hairpin resemble the binding site of L1 on the 23S rRNA [21]. The energy of formation of the secondary structures on Fig. 3 is –16.8 kcal/mol for *E. coli* and –12.2 kcal/mol for *H. influenzae*. Extension of the first hairpin of *E. coli* by additional base-pairing contributes another –8.1 kcal/mol. Finally, partial relaxation of the stem of the second (regulatory) hairpin, so that the Shine–Dalgarno box is downstream of the hairpin, is energetically favorable: in this case the energy is –20.0 kcal/mol for *E. coli* and –18.4 kcal/mol for *H. influenzae*.

Analogous structures regulate initiation of translation of this operon in other enterobacteria [22]. A more interesting case is that of archaea, where autoregulation of L1 is conserved even despite changes of the operon structure: in *Methanococcus vannielii* and *Halobacterium cutirubrum* this protein regulates initiation of translation in the operon *MvaL1* encoding the ribosomal proteins L1, L10 and L12 [23, 24]. For comparison, note that in *E. coli* and other

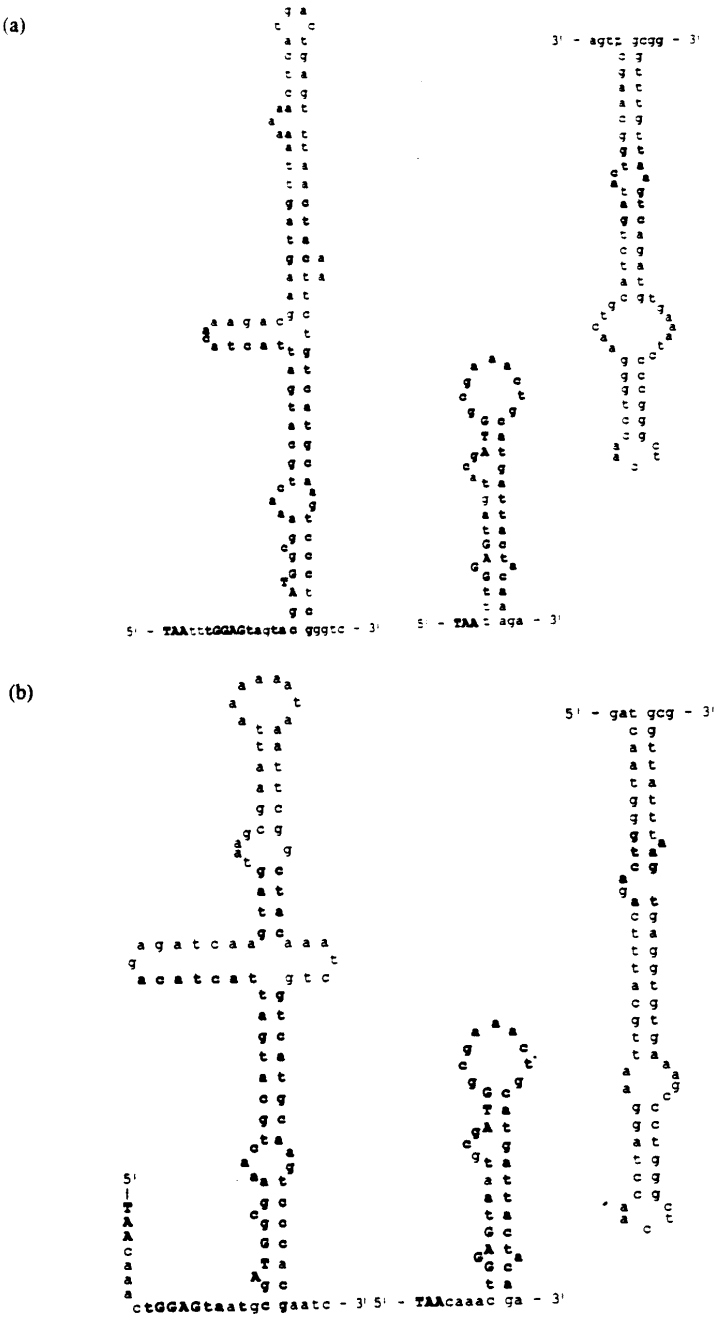


Fig. 1. RNA secondary structure in the regulatory region of the *spc* operon. (a) *Escherichia coli*. (b) *Haemophilus influenzae*. Left and center, two possible hairpins in the mRNA. Right, S8 binding site in the 16S rRNA. Capitals: the start codon ATG and the Shine–Dalgarno box GGAG of the gene *rplE*. Boldface: conserved nucleotides.

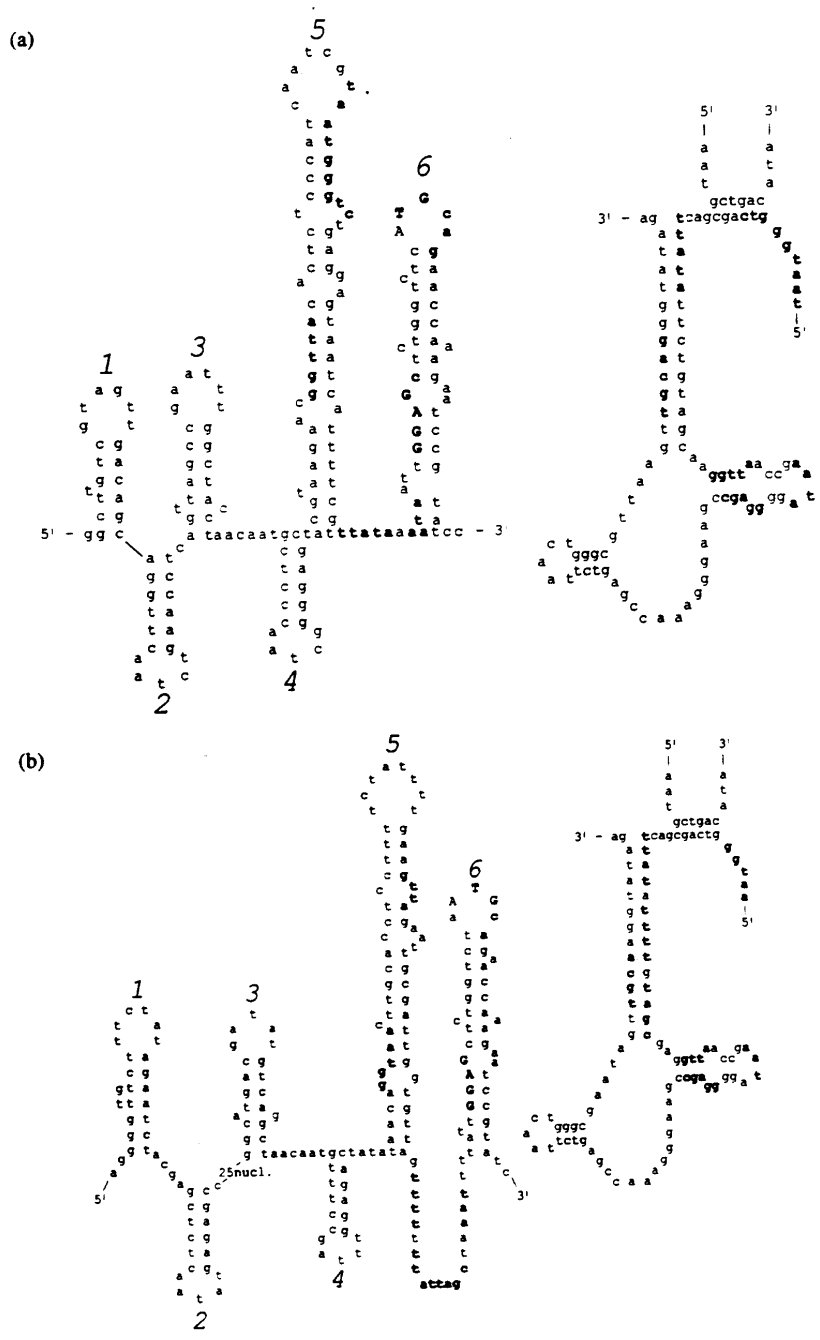


Fig. 2. RNA secondary structure in the regulatory region of the *S10* operon. (a) *Escherichia coli*, (b) *Haemophilus influenzae*. Left, mRNA secondary structure. Right, S10 binding site in the 23S rRNA. Capitals: the start codon ATG and the Shine–Dalgarno box GGAG of the gene *rpsJ*. Boldface: conserved nucleotides.

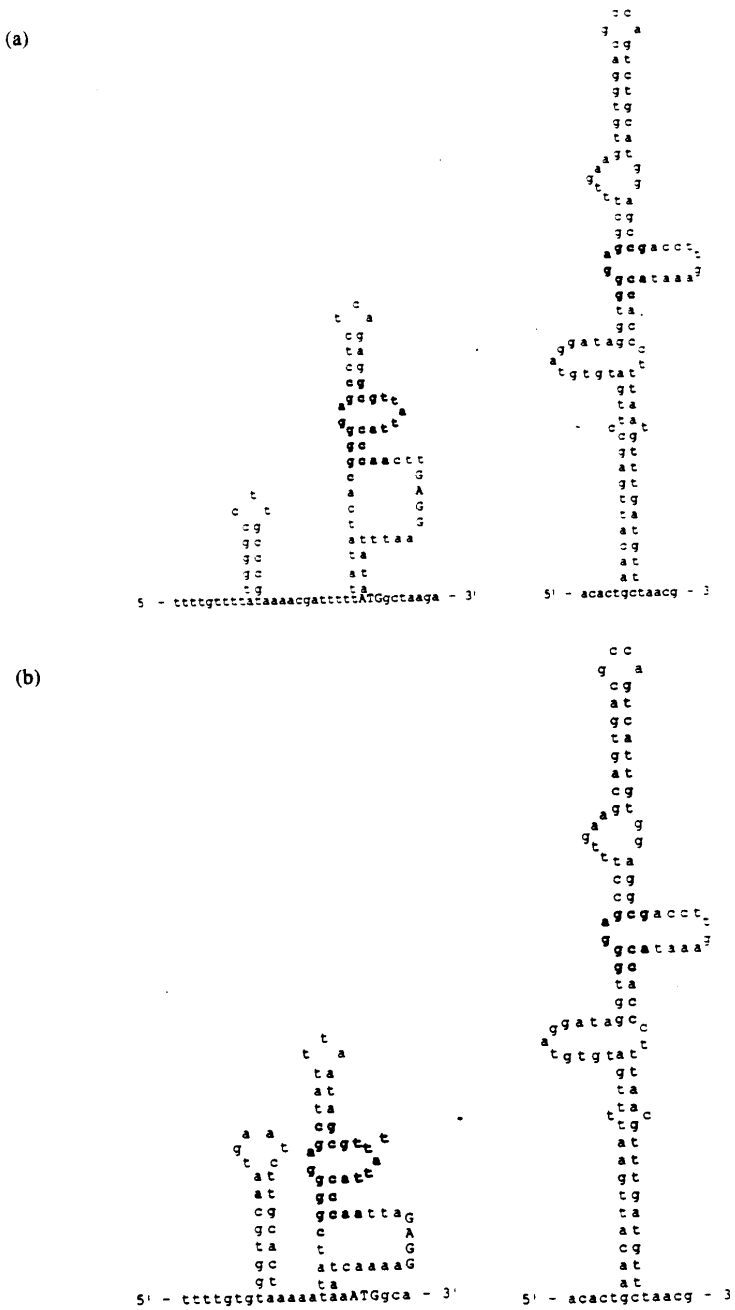


Fig. 3. RNA secondary structure in the regulatory region of the *L11* operon. (a) *Escherichia coli*. (b) *Haemophilus influenzae*. Left, mRNA secondary structure. Right, L1 binding site in the 23S rRNA. Capitals: the start codon ATG and the Shine–Dalgarno box GGAG of the gene *rplK*. Boldface: conserved nucleotides.

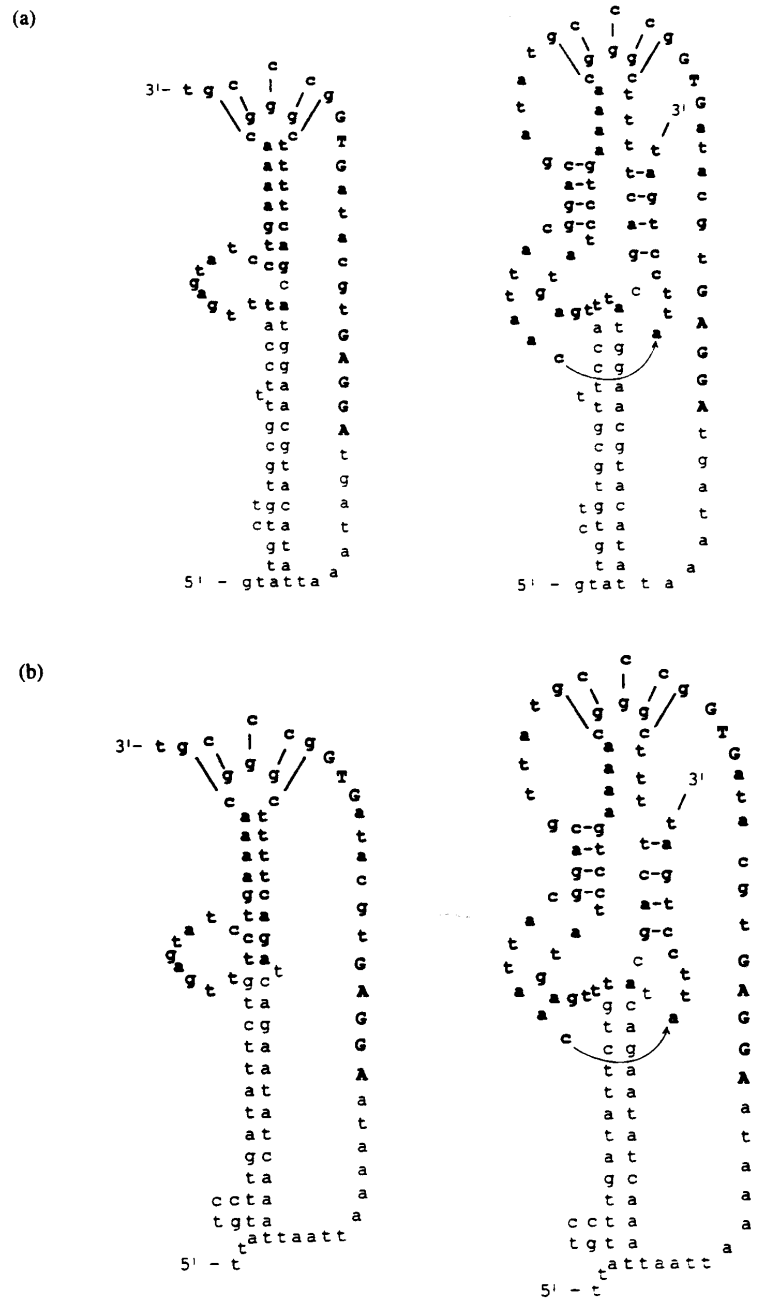
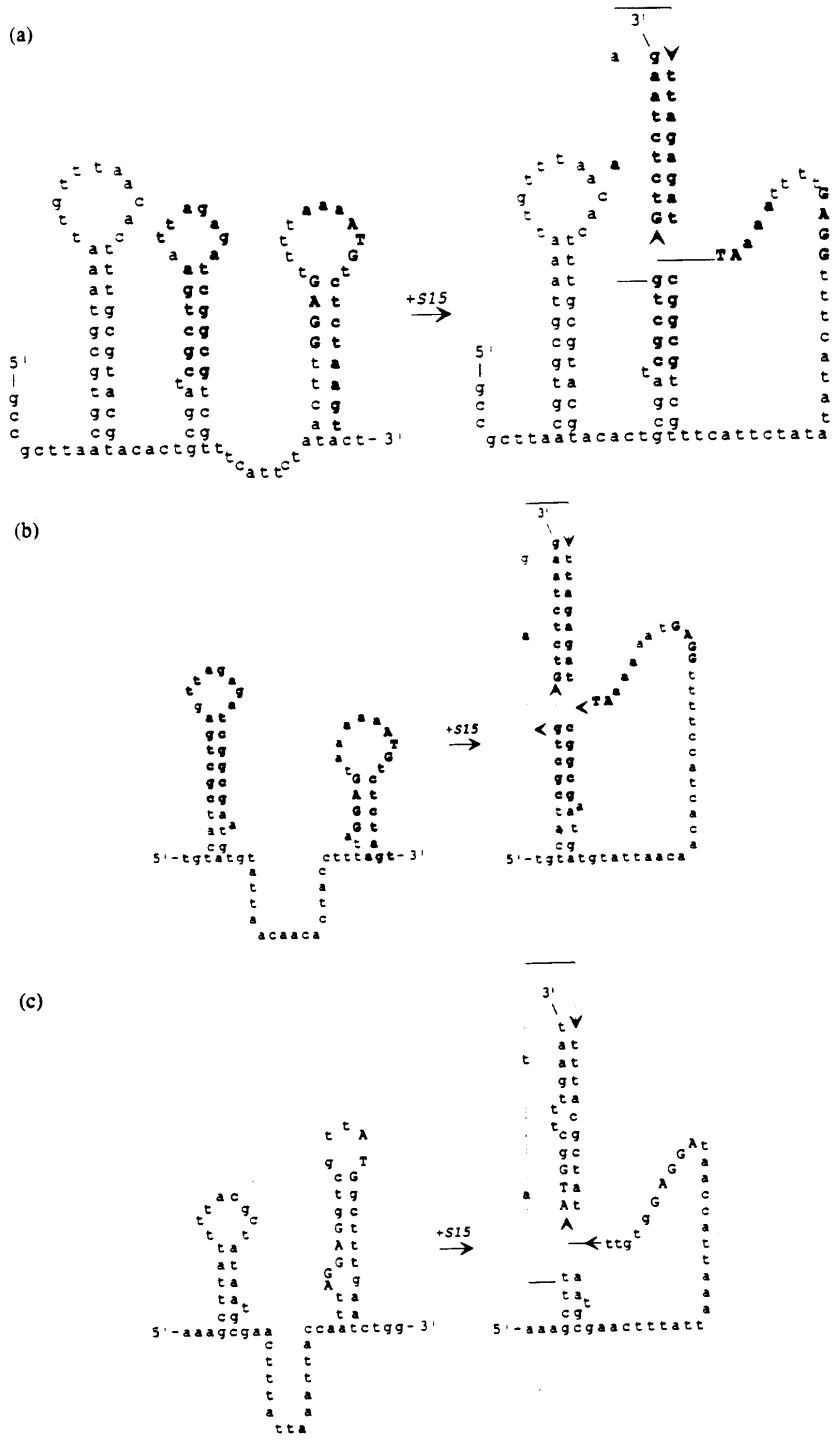


Fig. 4. RNA secondary structure in the regulatory region of the α operon. (a) *Escherichia coli*, (b) *Haemophilus influenzae*. Left, the bud of the pseudoknot recognized by S4. Right, the complete pseudoknot. Capitals: the start codon ATG and the Shine–Dalgarno box AGGAG of the gene *rpsM*. Boldface: conserved nucleotides.


(a)
+S15
5'-gcttaatacactgtttcattctata
(b)
+S15
5'-tgtatgtattaaca
(c)
+S15
5'-aaagcgaactttatt

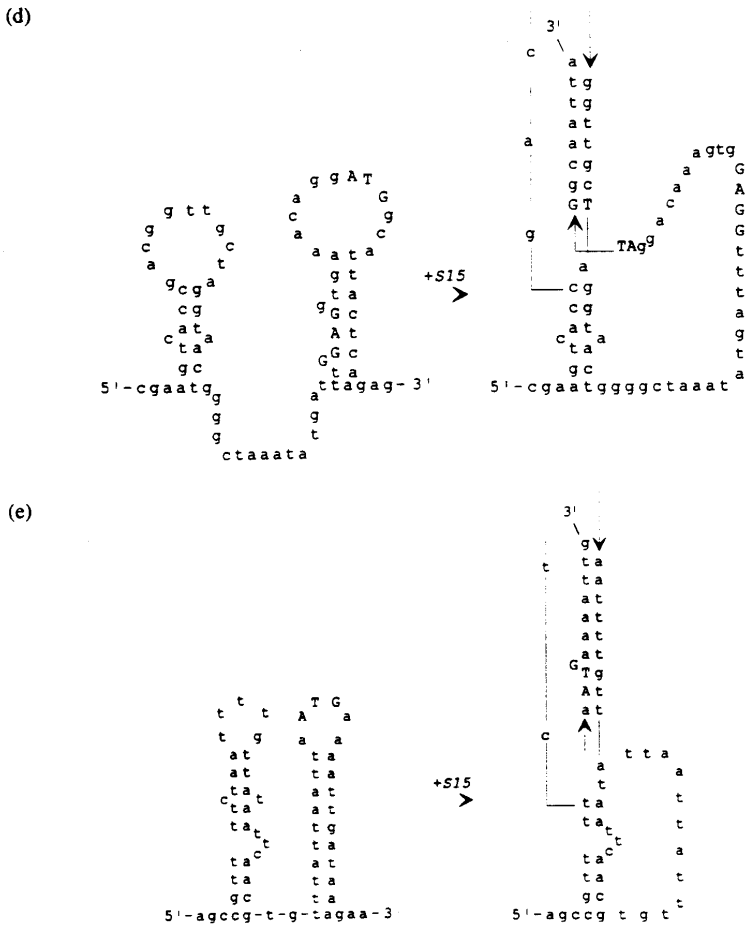


Fig. 5. Alternative RNA secondary structures in the regulatory region of the *S15* operon. (a) *Escherichia coli*, (b) *Haemophilus influenzae*, (c) *Helicobacter pylori*, (d) *Bacillus subtilis*, (e) *Mycoplasma genitalium*. Capitals: the start codon ATG and the Shine–Dalgarno box GGAG of the gene *S15*.

enterobacteria the genes *rplJ* and *rplL* coding for L10 and L12 respectively, form a separate group in the operon β autoregulated by the protein L10 [25, 26]. However, the location of these genes on the chromosome is the same in all cases, since in enterobacteria the operon β is immediately downstream the operon *rplKA*. Interestingly, in *H. cutirubrum* the sites recognized by L1 in the mRNA and in the 23S rRNA are similar, although they differ somewhat from the analogous sites of *E. coli* [24, Fig. 7]. Finally, the enterobacterial type of regulation of genes *rplKA* and *rplJL* has been suggested for an extremely thermophilic eubacterium *Thermotoga maritima* [27, 28].

Operon α. The regulatory site is a conserved pseudoknot with a complex configuration of secondary interactions (Fig. 4). This structure involved the Shine–Dalgarno box and the start codon of the first gene (*rpsM* coding for the protein S13). The pseudoknot is stabilized upon binding of the protein S4 encoded by the third gene of the operon [29].

Operon *S15* consists of a single gene. It is autoregulated by binding of S15 with one of the two alternative mRNA structures: a pair of hairpins or a pseudoknot (Fig. 5) [30]. These structures were observed not only in *E. coli* (–32.8 kcal/mol) and *H. influenzae* (–12.8 kcal/mol), but possibly in three other bacteria: *Helicobacter pylori* (–5.9 kcal/mol), *Bacillus subtilis* (–5.3 kcal/mol), and *Mycoplasma genitalium* (–6.6 kcal/mol). In the latter three cases both the hairpins and the pseudoknots are weaker than in the former two cases. *E. coli* has an additional hairpin. The Shine–Dalgarno box in all cases is either in the stem of the hairpin, or in the loop of the pseudoknot; the start codon is either in the loop of the hairpin, or in the stem of the pseudoknot.

Thus, for five operons of the ribosomal proteins of *H. influenzae* we have serious reasons to believe that the regulatory mechanism is the same as in *E. coli*. The common feature of all regulatory sites is the fact that they either imitate the binding site of the respective proteins in the ribosomal RNAs, or are pseudoknots. Often the regulation involves formation of alternative RNA structures. In all cases the start codon and/or the Shine–Dalgarno box are covered by the structural elements thus directly interfering with recognition of gene starts by the ribosomes. The pseudoknots formed by the α and *S15* mRNAs are pseudoknots of type H formed by two hairpins. This is the prevalent type of pseudoknots, probably because of the stabilizing influence of the helix stacking [31]. Interestingly, in all cases the start codon is positioned on the helix boundary and possibly is the site where the helical melting starts.

At the same time, construction of analogous structures could not be done for a number of operons (*str*, β, *rpmIrplT*, *rpsT*, *rpsA* regulated respectively by S7, L10, L25, S10, S1). As the evolutionary distance increases, the situation becomes even less uniform. Autoregulation of S15 seems to be conserved in *Helicobacter pylori* and Gram-positive bacteria *Bacillus subtilis* and *Mycoplasma genitalium*. A theoretical possibility for autoregulation has been demonstrated for L10 in *Thermotoga maritima* and for L1, not only in *T. maritima*, but in archaea *V. vannielii* and *H. cutirubrum*.

This situation resembles the relationships of transcription regulatory patterns. Homologous regulons of *E. coli* and *H. influenzae* have a conserved core, and the regulation may persist even despite changes of the operon structure. Sometimes partial conservation of regulons is observed even in distant comparisons, e.g. *E. coli* and *Bacillus subtilis*. At the same time, many regulons disintegrate, the main regulatory mechanism can change, or some genes may leave the regulon [32].

ACKNOWLEDGMENT

This study was partially supported by the Russian Foundation for Basic Research.

REFERENCES

1. Keener, J. and Nomura, M., *Escherichia coli and Salmonella. Cellular and Molecular Biology*, Neidhardt, F.C., Ed., Washington DC: ASM Press, 1996, vol. 1, Ch. 90, pp. 1417–1431.
2. Draper, D.E., *Trends Biochem. Sci.*, 1989, vol. 14, pp. 335–338.
3. Lindahl, L. and Zengel, J.M., *Annu. Rev. Genet.*, 1986, vol. 20, pp. 297–326.
4. Waterman, M.S., *Mathematical Methods for DNA Sequences*, Waterman, M.S., Ed., Boca Raton: CRC Press, 1989, Ch. 8, pp. 185–224.
5. Levitt, M., *Nature*, 1969, vol. 224, pp. 759–763.
6. Zuker, M., *Mathematical Methods for DNA Sequences*, Waterman, M.S., Ed., Boca Raton: CRC Press, 1989, Ch. 7, pp. 159–184.
7. Larsen, N., *Proc. Natl. Acad. Sci. USA*, 1992, vol. 89, pp. 5044–5048.
8. Bansal, A.K., Bork, P., and Stuckey, P.J., *Mathematical Modelling and Scientific Computing*, 1998, vol. 9, p. 123.
9. Fleischmann, R.D., *et al.*, *Science*, 1995, vol. 269, pp. 496–512.
10. Tomb, J.-F., *et al.*, *Nature*, 1997, vol. 388, pp. 539–547.
11. Kunst, F., *et al.*, *Nature*, 1997, vol. 390, pp. 249–256.
12. Frazer, C.M., *et al.*, *Science*, 1995, vol. 270, pp. 397–403.
13. Titov, I.I. and Ivanisenko, V.A., *Proc. 1st Int. Conf. on Bioinformatics of Genome Regulation and Structure*, Novosibirsk, 1998, vol. 2, p. 298.
14. Walter, A.E., Turner, T.D., Kim, J., Lyttle, M.H., Muller, P., Mathews, D.H., and Zuker, M., *Proc. Natl. Acad. Sci. USA*, 1994, vol. 91, pp. 9218–9222.

15. Cerretti, D.P., Mattheakis, L.C., Kearney, K.R., Vu, L., and Nomura, M., *J. Mol. Biol.*, 1988, vol. 204, pp. 309–329.

16. Henkin, T.M., Moon, S.H., Mattheakis, L.C., and Nomura, M., *Nucleic Acids Res.*, 1989, vol. 17, pp. 7469–7486.

17. Zengel, J.M. and Lindahl, L., *Biochimie*, 1991, vol. 73, pp. 719–727.

18. Zengel, J.M. and Lindahl, L., *J. Mol. Biol.*, 1990, vol. 213, pp. 67–78.

19. Shen, P., Zengel, J.M., and Lindahl, L., *Nucleic Acids Res.*, 1988, vol. 16, pp. 8905–8924.

20. Thomas, M.S. and Nomura, M., *Nucleic Acids Res.*, 1987, vol. 15, pp. 3085–3096.

21. Said, B., Cole, J.R., and Nomura, M., *Nucleic Acids Res.*, 1988, vol. 16, pp. 10529–10545.

22. Sor, F. and Nomura, M., *Mol. Gen. Genet.*, 1987, vol. 210, pp. 52–59.

23. Hanner, M., Mayer, C., Köhrer, C., Golderer, G., Gröbner, P., and Piendl, W., *J. Bacteriol.*, 1994, vol. 176, pp. 409–418.

24. Shimmin, L.C. and Dennis, P., *EMBO J.*, 1989, vol. 8, pp. 1225–1235.

25. Climie, S.C. and Friesen, J.D., *J. Mol. Biol.*, 1987, vol. 198, pp. 371–381.

26. Zhyvoloup, A.N., Kroupskaya, I.V., Lyaskovsky, T.M., and Paton, E.B., *Biopolimery i kletka*, 1994, vol. 10, no. 1, p. 37.

27. Liao, D. and Dennis, P.P., *J. Biol. Chem.*, 1992, vol. 267, pp. 22787–22797.

28. Paton, E.B. and Zhyvoloup, A.N., *Genetics*, vol. 32, no. 1, pp. 140–145.

29. Tang, C.K. and Draper, D.E., *Biochemistry*, 1990, vol. 29, pp. 4434–4439.

30. Philippe, C., Eyermann, F., Bonard, L., Portier, C., Ehresmann, B., and Ehresmann, C., *Proc. Natl. Acad. Sci. USA*, 1993, vol. 90, pp. 4349–4398.

31. Kolchanov, N.A., Titov, I.I., Vlasova, I.E., and Vlasov, V.V., *Progr. Nucl. Acids Res.*, 1996, vol. 53, pp. 131–196.

32. Gelfand, M.S. and Mironov, A.A., *Proc. 1st. Int. Conf. on Bioinformatics of Genome Regulation and Structure*, Novosibirsk, 1998, pp. 147–149.